

## **Синтаксический анализ в системе РМЛ**

Цель синтаксического анализа — построение синтаксических групп на одном морфологическом варианте одной клаузы. Группы строятся с помощью синтаксических правил ("форматок").

Клауза (фрагмент) – это простое предложение в составе сложного. Клауза может вкладываться в другую клаузу. Тогда она является подклаузой. С клаузой всегда связан некоторый набор вариантов вершины клаузы, где вершина клаузы - это некоторый омоним возможного слова данной клаузы. Подклаузы могут входить в синтаксические группы наравне с простыми словами.

Морфологический вариант – это набор **юнитов**, идущих друг за другом в данной клаузе. Юнит – это либо морфологический омоним одного слова, либо вариант подклаузы данной клаузы. Например, для фразы

*Дом, который построил Джек, развалился*

В главной клаузе, совпадающей со всем предложением, есть подклауза "который построил Джек". Морфологический вариант главной клаузы всегда будет состоять из четырех юнитов:

"Дом";

", который построил Джек";

", ";

"развалился".

Количество морфологических вариантов данной клаузы равно  $P_1 * P_2 * \dots * P_n$ , где  $n$  – число юнитов, а  $P_i$  – число разных вариантов юнита с номером  $i$ .

# Синтаксическая группа определяется следующими параметрами

номер первого и последнего юнита

тип группы (строка)

главная подгруппа

граммемы группы

В частном случае группа – один юнит, тогда номер первого и последнего юнита совпадают. Такую группу называют атомарной. Неатомарные группы подчиняются принципу проективности:

*Из того, что две группы пересекаются, следует, что одна лежит в другой (т.е. является ее подотрезком).*

Тип групп - это строковая константа (ПРИЛ\_СУЩ, ПГ и т.д.). Обычно одно синтаксическое правило строит один тип групп, но бывают исключения. Тип групп определяет во многом внутреннее устройство группы. Например, в группе "ПГ" главная подгруппа – предлог и т.д.

Главная подгруппа – это группа, которая является главной в данной группе. "Главность" – это понятие используется так же, как в теории HPSG (Head-Driven Phrase Structure Grammar). Например, для ПРИЛ-СУЩ главная группа – существительное.

Рекурсивно определяется понятие главного слова группы

Главное слово атомарной группы – само это единственное слово

Главное слово группы - это главное слово главной подгруппы



С помощью понятия "главной группы", например, определяется "именная группа": именная группа – это группа, у которой главное слово существительное.

Граммемы группы, или т.н. "внешние" граммемы группы, - это морфологические характеристики, которые должны определять поведение и сочетаемость группы во "внешнем" для нее мире, т.е. в группах, куда она сама входит. Внешние граммемы противопоставляются "внутренним" граммемам группы.

"Внутренние" граммемы – это граммемы главного слова группы. Например, во внешних граммемах группы "Петя и Вася" есть множественное число, а во внутренних его нет (главная группа однородного ряда – первая подгруппа, т.е. "Петя"). Например, в группе "два мальчика" внешние граммемы содержат именительный падеж, а главным словом является словоформа "мальчика" (без им. падежа).

# **Синтаксическое правило**

Синтаксическое правило получает на входе номер слова. От этого слова правило пытается построить новую группу определенного типа, соблюдая принцип проективности. На данный момент все правила пытаются объединить входную группу только с группами, находящимися от нее справа; в нашей записи правил (см. ниже) входная группа - это первая группа цепочки.

Все правила упорядочены (см. список ниже), поэтому в данной версии синтаксическая омонимия игнорируется, т.е. строится всегда только один вариант. Классический пример *древние стены города* наш анализатор разберет следующим образом: **генит\_иг( прил\_сущ(древние, стены ), города )**. Это происходит из-за того, что правило соединения прилагательного, согласованного с существительным, идет до правила, которое собирает генитивные цепочки.

Синтаксическое правило оперирует ограниченным числом объектов. Можно сказать, что они написаны на определенном подязыке C++. Основными объектами являются



Уже построенный набор групп,  
к которому нужно добавить  
новую группу



Характеристики отдельных  
слов – омонимов входного  
отрезка текста

Таким образом, синтаксические правила находятся выше омонимии, точнее омонимия находится вне этих правил. Рабочее название синтаксического правила – форматка.

## **Алгоритм работы**

На вход алгоритма подается морфологический вариант клаузы. В программе задан некоторый порядок применения правил. Этот порядок соответствует порядку построения групп: от меньших к большим. Например, сначала надо построить группы МОДИФ-ПРИЛ, а потом ПРИЛ-СУЩ, чтобы построить структуру на отрезку "очень красивый человек":

ПРИЛ-СУЩ (МОДИФ-ПРИЛ (*очень красивый*), человек )

Каждое правило применяется к каждому слову входного отрезка слева направо. Каждое правило для вновь построенной группы указывает ее главную группу, список грамем (обычно берется из списка грамем главного слова), тип.

# **Синтаксические правила**

Структура записи правил.

Запись правила состоит из следующих полей:

**Что**

в этом поле указывается цепочка групп (в частном случае слов), которая объединяется в новую синтаксическую группу.

**Условие**

в этом поле указываются условия, которые накладываются на компоненты новой группы.

**Главная группа**

главная группа, тип и граммемы

**Тип**

новой синтаксической

**Граммемы**

группы.

В скобках идет сокращенное название этого правила, которое совпадает с названием группы, если в результате работы правила построилась группа.